Investigating methods for using genetic data to estimate the association between host factors and being an infection source with application to probing the association between HIV infection and tuberculosis transmission

Isaac Goldstein[1], Damon Bayer[1], Ivan Barilar[2], Balladiah Kizito[3], Ogopotse Matsiri[3], Chawangwa Modongo[3], Nicola M. Zetola[4], Stefan Niemann[2*], Vladimir N. Minin[1*], Sanghyuk S. Shin[5*],

1 Department of Statistics, University of California, Irvine, California
2 German Center for Infection Research, Research Center Borstel, Germany
3 Victus Global Botswana Organisation, Gaborone, Botswana
4 Augusta University, Augusta, Georgia
5 Sue & Bill Gross School of Nursing, University of California, Irvine, California

* Co-corresponding authors: stniemann@yahoo.de, vminin@uci.edu, ssshin2@hs.uci.edu

Abstract

Identifying host factors that influence infectious disease transmission is an important step toward developing interventions to reduce disease incidence. Recent advances in methods for reconstructing infectious disease transmission events using pathogen genomic and epidemiological data open the door for investigation of host factors that affect onward transmission. While most transmission reconstruction methods are designed to work with densely sampled outbreaks, these methods are making their way into surveillance studies, where the fraction of sampled cases with sequenced pathogens could be relatively low. Surveillance studies that use transmission event reconstruction then use the reconstructed events as response variables (i.e., infection source status of each sampled case) and use host characteristics as predictors (e.g., presence of HIV infection) in regression models. We use simulations to study estimation of the effect of a host factor on probability of being an infection source via this multi-step inferential procedure. Using `TransPhylo` — a widely-used method for Bayesian estimation of infectious disease transmission events — and logistic regression, we find that low sensitivity of identifying infection sources leads to dilution of the signal, biasing logistic regression coefficients toward zero. We show that increasing the proportion of sampled cases improves sensitivity and estimation of logistic regression coefficients. Application of these approaches to real world data from a population-based TB study in Botswana fails to detect an association between HIV infection and probability of being a TB infection source. We conclude that application of a pipeline, where one first uses `TransPhylo` and sparsely sampled surveillance data to infer transmission events and then estimates effects of host characteristics on probabilities of these events, should be accompanied by a realistic simulation study to better understand biases stemming from imprecise transmission event inference.

Introduction

Better understanding of risk factors for transmitting infectious disease can help improve public health interventions. For example, HIV infection is associated with reduced

bacterial load and lower probability of cavitary disease among people with tuberculosis (TB) [3]. This has led to the hypothesis that HIV infection may reduce the probability of onward transmission of *M. tuberculosis* among people with TB. [3,4] Improved understanding of the effect of HIV on *M. tuberculosis* transmission could inform prioritization of scarce public health resources for contact tracing. Moreover, identifying host risk factors for transmission could provide valuable insights into the pathophysiology of the infectious disease.

In practice, it is rare to have perfect knowledge of who infected whom in an infectious disease outbreak, making study of risk factors for transmission difficult. Advances in methods for pathogen genomics have led to enhanced understanding of infectious disease transmission. [5,6] Sequencing data can be compared between infected hosts to rule out direct transmission when substantial genomic differences exist. More recently, several methods have been developed for using genomic and epidemiology data to reconstruct transmission events and identify a putative infection source for each transmission event [13,22–28]. In general, these tools were originally designed for outbreak investigations, where capturing most, if not all, infection recipients was feasible. Recently, researchers have used such methods on data collected in the context of passive disease surveillance as part of a statistical pipeline to determine the association between potential risk factors and being an infection source (transmitting an infection to at least one other host) [2,29,30]. **Fig 1** illustrates this pipeline. Researchers first infer a timed phylogenetic tree from genetic data, use the phylogenetic tree to infer infection source status labels for study participants, and finally infer the odds ratio of interest using the inferred infection source labels as observed response variables.

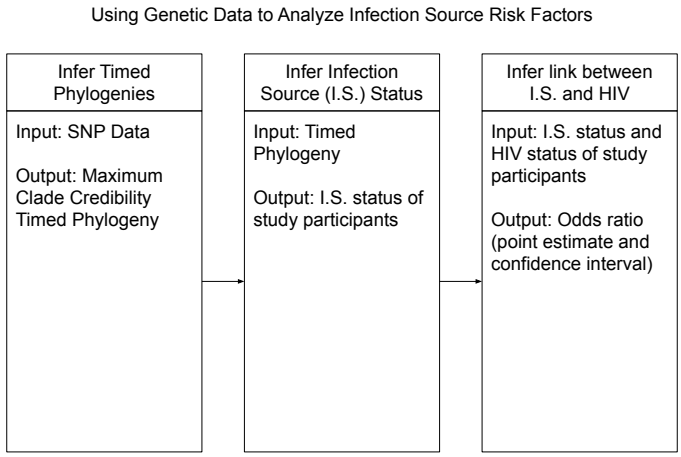

Fig 1. Visualization of statistical pipeline evaluated in this paper. Genomic data are used to construct timed phylogenetic trees, which are then used as inputs to infer transmission trees. Based on transmission trees, probability of being an infection source is derived, and then used as a response variable in an analysis of host factor associations.

While passive disease surveillance could capture significant portion of infectious

disease incidents, we expect a meaningfully smaller proportion of infection recipients will be sampled as compared to an outbreak investigation study. The performance of many of the transmission reconstruction methods in this somewhat different sampling context is not at all clear. And while statistical pipelines utilizing these transmission reconstruction methods have been used in practice, their ability to provide useful estimates of the associations between being an infection source and risk factors of interests remains largely untested. The goal of this study is to better understand the use of these pipelines in disease surveillance contexts by evaluating the performance of a set of statistical pipelines which use the commonly used `Transphylo` software to reconstruct transmission events [13]. We evaluate bias and precision of the estimators of odds ratios for probability of being an infection source when the response variable has been generated from SNPs via `TransPhylo`. We also evaluate the power of statistical tests based on these estimators to reject a null hypothesis that odds ratio is one, and coverage of corresponding 95% confidence and credible intervals. In particular, we evaluate the properties of using `TransPhylo` generated response variables and logistic regression, and `TransPhylo` generated response variables with two different statistical models that take into account measurement error. We accomplish this goal by simulating TB outbreaks with two classes of hosts (with and without HIV) with different probabilities of being an infection source where the underlying true odds ratios are known, and evaluating model performance under a variety of true odds ratios and sampling designs. Finally, we apply our pipelines to real data to investigate the association between HIV infection and being an infection source for *M. tuberculosis* transmission.

Our results show that performance of these pipelines varies dramatically depending on the true odds ratio. For modest effect sizes, power to reject a null hypothesis that the odds ratio is one can be low. Measurement error models often increase bias and confidence interval width past the point of usefulness. However, increasing proportion of cases sampled, duration of sampling period, or total sample size may improve pipeline performance. Overall, our findings suggest these pipelines should be used with caution, as their performance will depend highly on situational context.

Materials and methods

`TransPhylo` Overview

`TransPhylo` is an R package which uses a statistical model and Markov Chain Monte Carlo to produce a posterior distribution of transmission trees, graphs which describe who infected whom in an infectious disease outbreak. `TransPhylo` uses timed phylogenetic trees of sampled infection recipients in order to produce this posterior. One advantage of using `TransPhylo` for this task is that it is designed to allow for within-host evolution. It does this by "coloring" branches of the phylogenetic tree, where colors represent hosts, and allowing for color changes to occur anywhere on a branch, as opposed to just at bifurcations of the tree. Another advantage of `TransPhylo` is that it explicitly allows for outbreaks where not all individuals have been sampled. Using the posterior distribution of transmission trees, we can create a probability of being an infectious source for each sampled individual by calculating the proportion of posterior trees in which the individual transmitted the disease to another individual (regardless of whether the disease recipient was sampled).

Simulation Set Up

All simulated data sets were simulated in R version 4.0.2 or 4.0.4 using the `nosoi` package version 1.0.3 [11,14]. `nosoi` simulates an infectious disease outbreak by

simulating individual infection recipients which generate new infection recipients in discrete units of time until either a specified time limit is reached or a specified total number of infection recipients is reached. The behavior of infection recipients is controlled through three parameters: the probability of the recipient ceasing to be infectious at the beginning of a time step, the number of susceptible persons an infection recipient encounters in a time step, and finally, the probability that, given a contact with a susceptible individual, an infection recipient will transmit the infection. For all simulations in this study, we assumed discrete time steps represented one month of real time, and chose model parameters to mimic a plausible *M. tuberculosis* outbreak (see Appendix A.1.1 for simulation parameters). In brief, R_0 was set to be 1.18, the mean latent period was 9 months among hosts that became infectious, and the mean infectious period was 3 months. All outbreaks used in this study lasted 8 years. In order to control the number of sampled recipients, we generated simulations at random, but only used outbreaks with a total number of between fifty and two thousand infection recipients. This meant that many potential outbreaks were rejected as part of our sampling simulation process. While this rejection sampling has the potential to change the true odds ratio being estimated, we have found empirically this is not an issue for outbreaks lasting 8 years. In our simulation set-up, each outbreak represents a cluster of *M. tuberculosis* infections to be analyzed jointly using `Transphylo`, mimicking the common real world practice of clustering SNPs by some threshold and generating separate timed phylogenies for each cluster. Henceforth we will refer to these outbreaks as clusters.

We simulated clusters with two types of individuals representing people living with and without HIV co-infected with *M. tuberculosis*. Based on data collected for a population-based epidemiology study of *M. tuberculosis* in Botswana (discussed below), we specified that on average 53 percent of infection recipients would be among people living with HIV. We had five primary simulation settings where the only changing parameter was the ratio of the probabilities of transmission given contact for two classes of infection recipients. The settings were: probability of transmission given contact was three times lower among people living with HIV compared to that among people without HIV, 1.75 times lower, equal, 1.75 times higher, or ten thirds higher. All other parameters were identical for the two classes of infection recipients.

For each simulation setting, we simulated one hundred sets of fifty clusters lasting eight years where recipients were sampled in the last three years of the outbreak duration. Recipients were only sampled during their infectious periods, with the sampling time equally likely at any point in this time frame. For each cluster, we sampled 16% of eligible recipients. Only simulations where all fifty clusters had at least two sampled individuals were accepted. We chose this sampling scheme to reflect real world sampling conditions in our study setting, where data were collected as part of ongoing surveillance of an endemic disease, and where there was little chance of sampling recipients during their latent period. For example, due to under-diagnosis, the World Health Organization estimated that only 54% of people with new TB episodes in Botswana were reported in 2019 [7]. Using this proportion, we divided the number of *M. tuberculosis* sequences available by the number of estimated TB episodes (reported + unreported) to calculate the sampling proportion for our study. Thus, 16% reflects a reasonable sampling proportion one would expect when data come from passive surveillance of TB.

Once sampled, we generated timed phylogenetic trees for each cluster using `nosoi`, the trees from separate clusters represent the trees from separate transmission clusters which would be generated as part of a real data analysis. These timed phylogenetic trees served as the input for `TransPhylo` (version 1.4.4).

In addition to our primary simulations, we conducted a number of secondary simulations to further investigate the operating characteristics of our statistical pipeline.

For all secondary simulations, people living with HIV had a probability of transmission given contact that was set to be 1.75 times lower than people without HIV. We then varied the sampling proportion, increased total sample size, and decreased time from start of epidemic to start of sampling. To investigate the impact of increasing the sampling density we increased the sampling proportion from 16% to 32% while keeping the sample size approximately the same by halving the number of clusters in a simulation (from fifty to twenty-five). To investigate the impact of an increase in sample size, in the second setting we kept the sampling proportion at 16%, but doubled the number of clusters from fifty to one hundred. To investigate the impact of shortening the time from beginning of epidemic to beginning of sampling, in the third setting, we kept the sampling proportion at 16%, but increased the sampling window from the last three years to the last seven years of the simulation, while decreasing the number of clusters from fifty to thirty-five to approximately maintain sample size.

Using `TransPhylo` on Simulated Data Sets

We used the multiple cluster functionality of `TransPhylo` to analyze all simulated trees (from multiple simulated clusters) at the same time, where all disease parameters were shared across all trees, while the sampling proportion π was allowed to vary within each tree. Based on empirical results from early simulations, for the primary simulations, the sampling prior was set as a beta distribution with alpha equal to 1 and beta equal to 19. The other priors in `TransPhylo` cannot be changed by the user. `Transphylo` assumes the number of individuals an infectious person will infect has a negative binomial distribution governed by two parameters r and p. These parameters also describe the reproduction number, and so their priors control the prior for the reproduction number. r has an exponential prior with rate parameter one, p has a uniform prior over the interval zero to one. We found in practice that p was not identifiable and fixed it at 0.5 for our analyses. Finally, the within-host effective population size also has an exponential prior with rate parameter one. The generation time and sampling time distribution were identical, with shape parameter 10, scale parameter 0.1. Under some assumptions, this distribution matches the mean of the generation time specified in our simulations [34]. We sampled 100,000 draws from the posterior distribution of transmission trees, discarding the first 50,000 as burn-in, and saving only every tenth sample. Probability of being an infection source was calculated as the proportion of posterior draws in which a disease recipient transmitted the disease. We considered an infection recipient to be an infection source if their inferred posterior probability of being an infection source was greater than 0.6. We have found the exact value of this cutoff makes little difference, as the distribution of probabilities tends to be extremely bi-modal, clustered around one and zero (see appendix A.5).

Inferring Odds Ratios using Infection Source Labels

All statistical inference was conducted in R (either version 4.0.2 or 4.0.4). We estimated odds ratios through logistic regression via the `glm` function in R. The mean of the response variable was the probability of being an infection source and the covariate was HIV status, where one indicates a person living with HIV, and zero indicates a person without HIV. That is:

$$\log \frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + HIV\beta_1,$$

where $Y = 1$ when a infection recipient is labeled as an infection source and $Y = 0$ when they are not. For simulations, we used logistic regression using both the true infection source labels and the labels inferred by `TransPhylo`. True odds ratios for each

simulation setting were calculated using Monte-Carlo simulations (see appendix A.4). Our results suggested logistic regression using the `TransPhylo` inferred labels yielded sub-optimal inference due to mislabeled infection recipients (i.e. `TransPhylo` inferred an infection recipient was not an infection source when in fact they were). This is a well-known phenomenon in statistics called measurement error, see Neuhaus (1999) for a clear introduction to the problem [20]. Many methods have been developed to adjust for this misclassification, we chose to include two as potential variations to our statistical pipeline.

The first method is available in the `SAMBA` package by Beesley and Mukherjee (2020) (we used version 0.9.0) [15]. The `SAMBA` misclassification model assumes no misclassification of negative results (specificity is 1) and first infers the sensitivity of the response variable, then incorporates this sensitivity in the final logistic regression. To infer sensitivity, the user is required to input a fixed unconditional probability of the response variable being one, in our setting, this is the unconditioned probability that an infection recipient is an infection source. For simulations, the true unconditional probability from the sampled data was used. For the real data analysis, the mean probability of being an infection source in the last three years of the cluster from one of the primary simulation settings was used.

The second method was a Bayesian approach we created to account for the case when specificity is not 1. For this model we assume the sensitivity and specificity for the response labels are known. In the model below Y represents the observed response label whereas Z represents the unobserved true response label.

$$
\begin{aligned}
\beta_0 &\sim N(0,1),\\
\beta_1 &\sim N(0,1),\\
\log\frac{P(Z_i=1)}{1-P(Z_i=1)} &= \beta_0 + HIV_i\beta_1,\\
Y_i \mid Z_i = 0 &\sim \text{Bernoulli}(p = 1 - \text{Specificity}),\\
Y_i \mid Z_i = 1 &\sim \text{Bernoulli}(p = \text{Sensitivity}).
\end{aligned}
$$

For the simulations, Monte-Carlo estimated sensitivity and specificity were used, rounded to two significant digits. For the real data analysis, the Monte-Carlo estimated sensitivity and specificity from one of the primary simulations was used. There was little difference in estimated sensitivity and specificity between the different settings (**Figure A1**). We used Hamiltonian Monte Carlo (HMC) to conduct Bayesian inference with this model using Stan through the `rstan` package (version 2.21.2). We generated 2000 posterior draws and discarded the first 1000 as burn-in. We used the median of the posterior distribution for point estimates and 95% Bayesian credible intervals as measures of uncertainty.

Data Collection

The real data used in this study was originally collected as part of a molecular epidemiology study of *M. tuberculosis* transmission in two regions of Botswana. Botswana has a high prevalence of HIV as well as *M. tuberculosis*. In 2017 19.9% of 15-49 year were living with HIV, 300 of 100,000 people had TB, and 48% of people with TB also had HIV [21]. All patients who were diagnosed with *M. tuberculosis* between 2013 and 2016 in the two regions were eligible for enrollment in the study. Of the 4331 enrolled participants, 2159 had culture-confirmed TB. Any participants who had unknown HIV status or who had had a negative result more than a year prior to sampling were offered HIV testing as part of the study protocol [8,10].

Sequencing

Of the *M. tuberculosis* DNA extracted from cultured isolates, 1671 (77%) were of sufficient quality to undergo sequencing using Nextera XT libraries. We performed WGS using Illumina Technology (MiSeq, NextSeq 500) generating 2x151bp paired end reads. All genomes were analysed with the MTBseq pipeline using the standard input values [9]. In short, reads were mapped to the *M. tuberculosis* H37Rv genome (GenBank ID: `NC_000962.3`) using BWA-mem and Samtools. Base call recalibration and realignment of reads around insertions and deletions was performed using GATKv3. Finally, variant calling was performed with Samtools mpileup and in-house scripts that employed thresholds for coverage and base quality. After sequencing, 1426 libraries were of high enough quality, defined by at least 50x coverage and 95% of the reference genome covered, to be used for further analysis. Average coverage of the good quality samples was 120x.

Concatenated sequence alignment of variants was produced by including those genome positions that fulfilled the aforementioned criteria for coverage and variant frequency in 95% of all samples. Repetitive regions (PPE/PE-PGRS genes), InDels, consecutive SNPs in 12bp windows and genes implicated in antibiotic resistance were excluded. Phylogenetically informative SNPs were taken from existing literature. We excluded sequences with heterogeneous reads along a base position, which is indicative of mixed infections, leaving 1347 sequences for analysis.

Analysis of Real Data

Within each lineage, we calculated raw single nucleotide polymorphism (SNP) distances between all sequences and clustered sequences based on two cutoffs, either sequences with a SNP distance of five or fewer, or ten or fewer were clustered together. We analyzed only clusters with four or more sequences where there were at least two times the number of sequences in the cluster minus two varying sites amongst all sequences in the cluster. For the five SNP analysis, we had 147 sequences from 21 clusters. For the ten SNP analysis, we had 388 sequences from 46 clusters to analyze.

We used BEAST 2 version 2.6.3 to create posterior distributions of timed phylogenetic trees with 10 million posterior draws for each of our clusters [16]. We then calculated maximum clade credibility trees to use as the inputs into `TransPhylo`, discarding the first ten percent of posterior trees. We used an HKY model with strict molecular clock and constant population model, and used an ascertainment bias correction for sequences of unknown length to account for the fact that analyzed sequences included only varying sites [17]. See appendix A.1.2 for all prior specifications.

For `TransPhylo`, shape and scale parameters for the generation time and sampling time distributions were fixed at 10 and 0.1 respectively. The multitree function was used to conduct inference simultaneously on all clusters generated by the same SNP threshold, with all parameters but the sampling proportion shared amongst all clusters. We used a beta distribution with alpha set to 1 and beta set to 19 as a prior for the sampling distribution. We drew 100000 posterior samples, discarding the first 50000 samples as burn-in and saving only every tenth sample. We calculated the odds ratio for people living without HIV vs with HIV using the inferred `TransPhylo` labels using logistic regression, using the SAMBA package, and using the Bayesian model as described in a previous subsection.

Results

Simulated Data

Table 1 shows summary statistics of the simulated clusters used in the primary simulation setting where people living with HIV were 1.75 times less likely to transmit infection given contact compared to people without HIV. Median cluster size was 116, with 95% of clusters ranging from 53 to 393 hosts. Latent periods for infection recipients had a median of 9 months, while infectious periods lasted a median of 3 months. Median tree height was 5.9 years.

Metric	Median	[2.5%Q, 97.5% Q]
Cluster Size	116	[53, 393]
Sample Size	9	[3,34]
Latent Period (Months)	9	[4,17]
Infectious Period (Months)	3	[1,15]
$P(IS^*)$	0.69	[0.64, 0.73]
$P(HIV^+)$	0.53	[0.49,0.56]
$P(IS\|HIV^-)$	0.77	[0.71, 0.81]
$P(IS\|HIV^+)$	0.61	[0.57, 0.67]
$P(HIV^+\|IS = N)$	0.65	[0.58, 0.71]
$P(HIV^+\|IS = Y)$	0.48	[0.43, 0.52]
Tree Height (Years)	5.92	[2.59, 7.42]

*Infection Source

Table 1. Simulation Summary Statistics for 8 Year Simulation. Summary statistics of simulated infectious disease outbreaks from 100 simulations where probability of transmission given contact for hosts without HIV was 1.75 times larger than for hosts with HIV.

Operating Characteristics of Statistical Pipeline

For each simulation setting, we calculated the sensitivity and specificity of `TransPhylo` for identification of an individual as an infection source. Point estimates for each simulation setting are displayed in **Figure A1**. In general, specificity was high, around 0.97, while sensitivity was low, approximately 0.26. Sensitivity increased when either the sampling density or the sampling window was increased, to approximately 0.4. Then, for each version of our pipeline and each simulation setting, we calculated frequentist coverage, power to reject a null hypothesis that the odds ratio was one (with significance level 5%), percent bias (bias divided by the true odds ratio), and mean confidence interval width of 95% confidence intervals. **Figure 2** shows the results from the primary set of simulations. The four pipelines tested for each simulation were: a reference pipeline where true infection source labels were known and used as inputs into a generalized linear model (Truth + GLM), a pipeline where `TransPhylo` labels were used as input into a generalized linear model (TP + GLM), a pipeline where `TransPhylo` labels were used as input into SAMBA (TP + ME1) for misclassification adjustment and a pipeline where `TransPhylo` labels were used as inputs in our Bayesian model for misclassification adjustment (TP + ME2). Across simulation settings, the Truth + GLM pipeline had coverage around 0.95, power above 0.95 (except in the case when the odds ratio was 1), percent bias near 0 and confidence interval width ranging from 0.16 to 2.6. For the TP + GLM pipeline, coverage varied depending on the simulation setting, ranging from 0.52 to 0.93. When the true odds ratio was not one, TP + GLM had power ranging from 0.54 to 1. Percent bias ranged from -24% to 54%,

mean credible interval width was 0.326 to 2.38. The TP + ME1 pipeline had coverage and power similar to TP + GLM, with percent bias ranging from -54% to 198% and MCIW between 0.25 and 23.9. Finally, the TP + ME2 pipeline had coverage ranging from 0.74 to 0.97, power ranging from 0.42 to 1, percent bias from -26% to 121% and MCIW from 1.12 to 21.2.

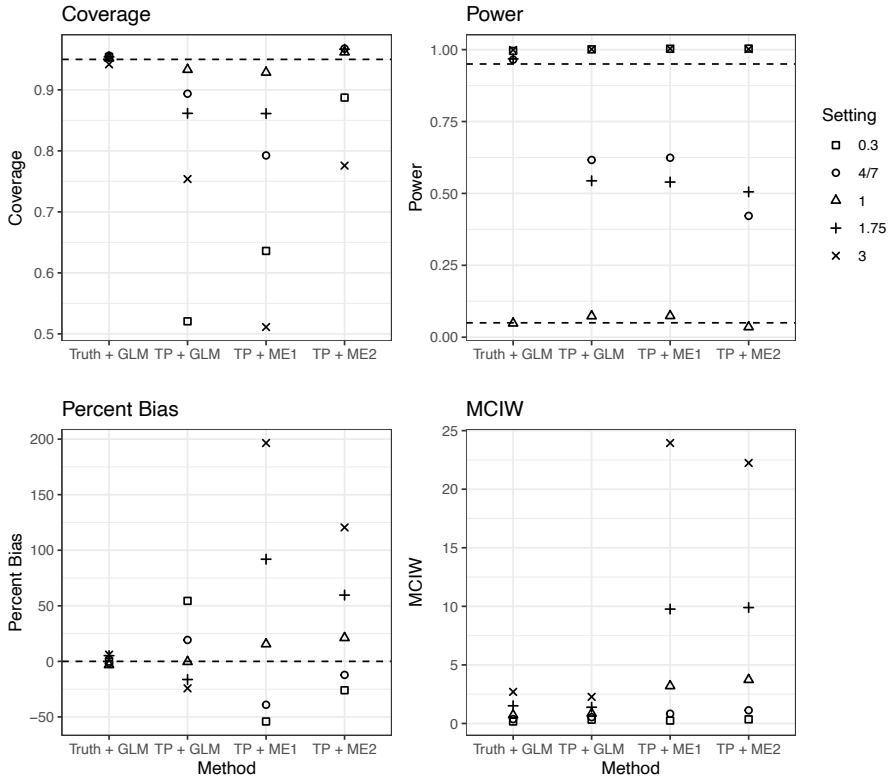

Fig 2. Operating characteristics of statistical pipelines in primary simulation settings. Truth + GLM is a reference pipeline where the true infection source labels are used as response variables in a generalized linear model. TP + GLM is a pipeline where infection source labels generated by `TransPhylo` are used as response variables in a generalized linear model. TP + ME1 is a pipeline where infection source labels are generated by `TransPhylo` and used as an input into a model from the `SAMBA` package, which allows for false positives. TP + ME2 again uses labels generated by `TransPhylo` as response variables into a model which allows for both false positives and false negatives. The settings denote different simulation settings, each value describes the true ratio of the probability of transmission given contact for infection recipients living without HIV to infection recipients with HIV. I.E., 4/7 means that the probability of transmission given contact for hosts without HIV was 1.75 times as small as the probability of transmission given contact for hosts with HIV. Coverage refers to proportion of simulations where 95% confidence intervals captured the true odds ratio. Power refers to the proportion of simulations where a null hypothesis that the true odds ratio was one would be rejected, assuming significance level of 5%. Percent bias is bias divided by the true odds ratio, MCIW is mean confidence interval width.

The results of our secondary set of simulations (along with the original results from the 1.75 setting) are displayed in **Figure 3**. Increasing the percent of hosts sampled from 16% to 32% (Increase Sampling Density) led to increased coverage from 0.86 to 0.9

for the TP + GLM pipeline as compared to the default 1.75 scenario, and increased power to reject a null hypothesis that the odds ratio was one from 0.54 to 0.8. Percent bias decreased from -16% to -15%, while MCIW decreased from 1.38 to 1.22. Increasing the sampling window to all but the first year of the outbreak led to decreased coverage of 0.85, increased power of 0.82, decreased percent bias of -10%, and decreased MCIW of 1.31. Doubling the number of clusters analyzed from fifty to one hundred (Increase Clusters) decreased coverage to 0.82, increased power to 0.94, decreased percent bias to -14%, and decreased MCIW to 0.98. For the TP+ME1 pipeline, the default coverage was 0.86, while coverage in each of the three alternate settings ranged from 0.86 (Increased Sampling Density) to 0.7 (Increase Clusters). Default power was 0.54, which increased in all three alternate settings, ranging from 0.8 (Increase Sampling Window) to 0.94 (Increase Clusters). Default percent bias was 91.9%, which increased to 96% or 98% for Increase Clusters and Increase Sampling Window respectively, but decreased to 61% for Increase Sampling Density. Default MCIW was 9.77, which decreased for all alternate simulation settings, ranging from 5.69 (Increase Sampling Density) to 7.4 (Increase Sampling Window). For the TP + ME2 pipeline, default coverage was 0.97, which decreased for all alternate scenarios, ranging from 0.74 (Increase Clusters) to 0.9 (Increase Sampling Density). Default power was 0.51, which increased for all alternate scenarios, ranging from 0.77 (Increase Sampling Density) to 0.95 (Increase Clusters). Default percent bias was 60%, which decreased to 59% for Increase Sampling Density, but increased to 99% and 100% for Increase Sampling Window and Increase Clusters respectively. Default MCIW was 10, which increased for Increase Sampling Window to 10.54, but decreased to 7.89 for Increase Sampling Density and decreased slightly to 9.9 for Increase Clusters.

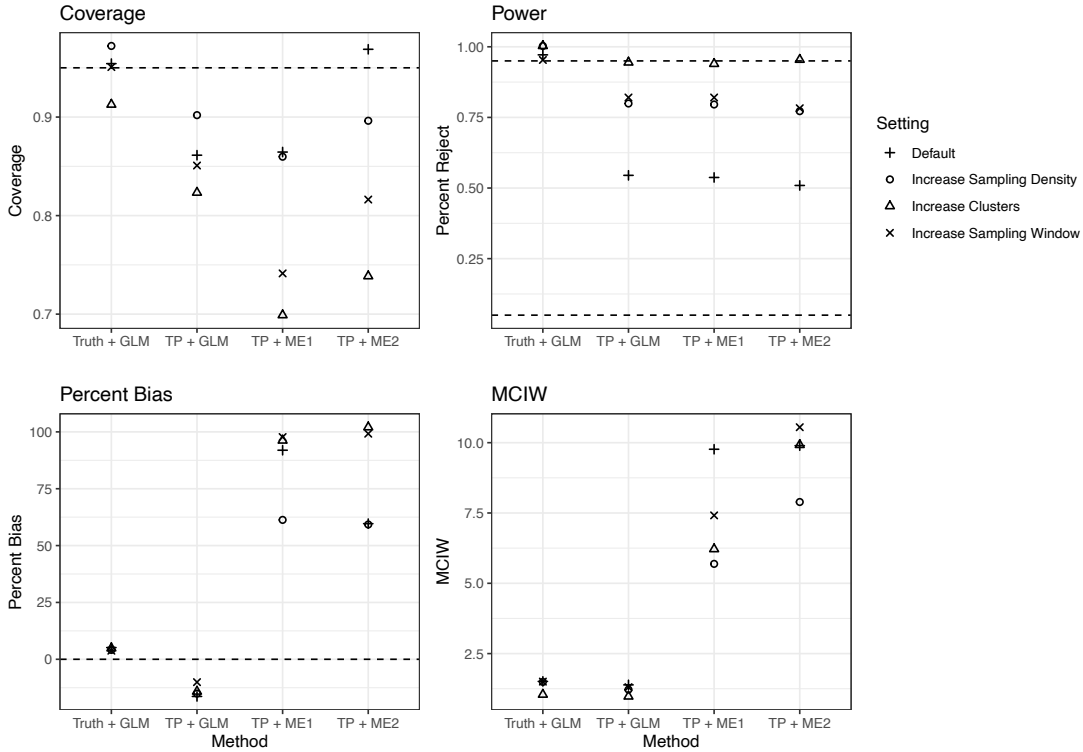

Fig 3. Operating characteristics of statistical pipelines in secondary simulation settings. Truth + GLM is a reference pipeline where the true infection source labels are used as response variables in a generalized linear model. TP + GLM is a pipeline where infection source labels generated by `TransPhylo` are used as response variables in a generalized linear model. TP + ME1 is a pipeline where infection source labels are generated by `TransPhylo` and used as an input into a model from the `SAMBA` package, which allows for false positives. TP + ME2 again uses labels generated by `TransPhylo` as response variables into a model we created which allows for both false positives and false negatives. The settings denote different simulation settings. For all simulations the probability of transmission given contact for infection recipients living without HIV was 1.75 times as large as the probability of transmission given contact for infection recipients living with HIV. Default refers to the simulation settings used in the primary simulation setting. In the Increase Sampling setting percent of active infection recipients sampled was doubled from 16% to 32%. In the Increase Clusters setting the number of clusters sampled from was doubled from 50 to 100. In the Increase Sampling Window the sampling window was changed from the last three years of the simulation to the last seven years. Coverage refers to proportion of simulations where 95% confidence intervals captured the true odds ratio. Power refers to the proportion of simulations where a null hypothesis that the true odds ratio was one would be rejected, assuming significance level of 5%. Percent bias is bias divided by the true odds ratio, MCIW is mean confidence interval width.

Real Data Results

Study participant characteristics from the Botswana study are summarized in **Table 2**. The vast majority of sampled TB bacteria belonged to lineage 4. More than half of study participants were male, and most participants were living with HIV.

	Count	Percent
TB Lineage		
Lineage 1	83	6.4
Lineage 2	69	5.3
Lineage 3	12	0.9
Lineage 4	1142	87.4
Sex		
Male	719	55.1
Female	587	44.9
Age		
Under 18	80	6.1
18 to 35	693	53.1
35 to 55	466	35.7
55 to 75	67	5.1
75 and older	10	0.8
HIV Status		
Without HIV	590	45.2
With HIV	716	54.8

Table 2. Study participant characteristics Summary statistics of participant characteristics from the study of *M. tuberculosis* in Botswana.

The four largest phylogenetic trees from each of our real data analyses (using five SNPs and ten SNPs as clustering cutoffs) produced by our statistical pipelines are displayed in **Figure 4**. From our genetic samples, we used BEAST2 to create a phylogenetic tree, then used `TransPhylo` to label tips of the tree as being infection sources or not infection sources and compare with their recorded HIV status. The bottom row have tips labeled by ID numbers, demonstrating the way trees change when inclusion criteria change. All participants from the five SNP tree (right) remain in the same ten SNP tree (left), with additional participants included because of the relaxed inclusion criteria.

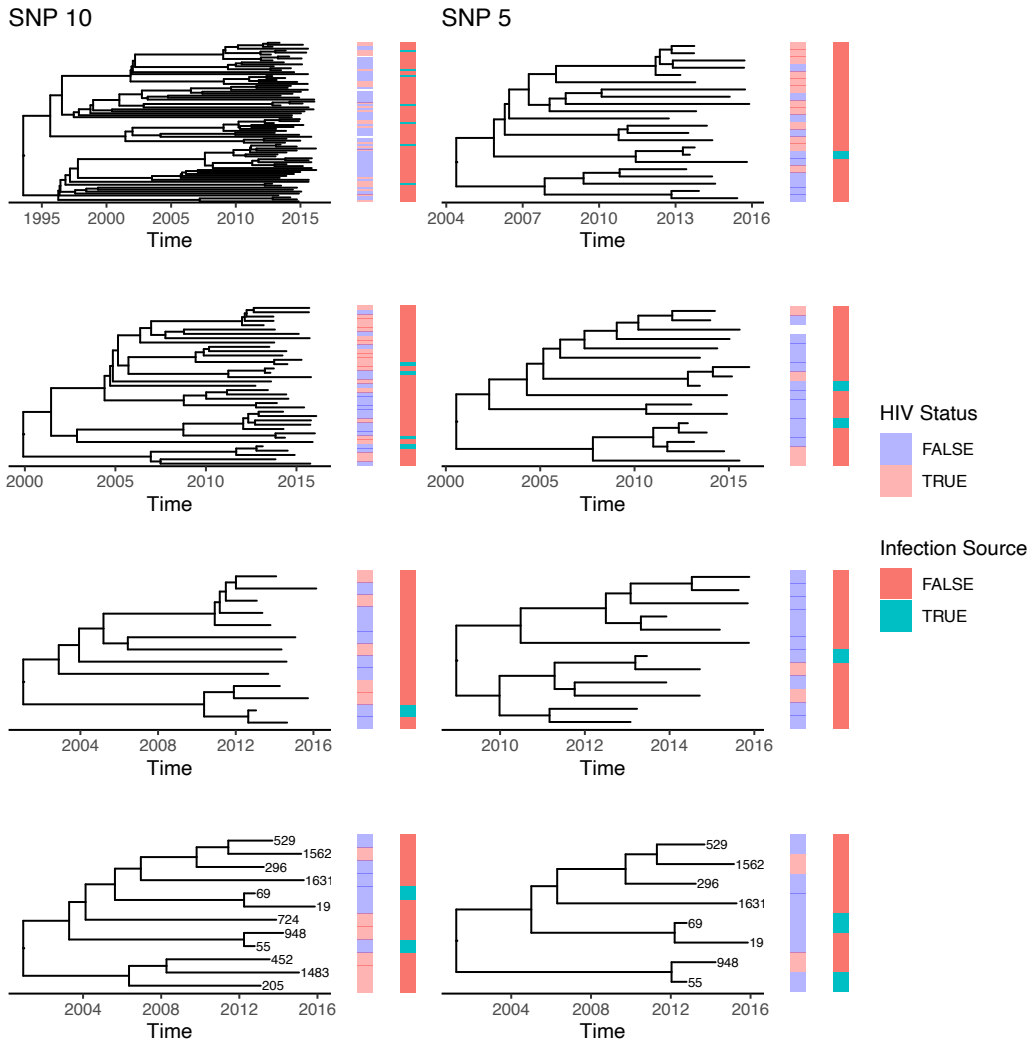

Fig 4. Timed Phylogeny with tips labeled by HIV status and inferred infection source status using `TransPhylo`. Timed Phylogenies are generated in `BEAST 2`. Before timed phylogenies are generated, TB genomes are first clustered by SNP distance based on a cutoff of either five or ten SNPs. The trees generated under each cutoff are different, but related to each other, with larger trees overall using a cutoff of ten SNPs as opposed to five.

Table 3 shows results from our real data analysis using clusters generated under a five SNP cutoff rule and a ten SNP cutoff rule. Mean cluster size and mean tree height increased marginally depending on which cutoff was used (increasing from 7 to 8.43 participants and from 6.86 years to 8.27 years), but both are similar to cluster size and tree height in simulated data (**Table 1**). Overall sample size was smaller than in simulation settings. In both analyses, between 11% to 12% of participants were identified as infection sources. Using any iteration of our pipeline, we fail to reject (at a 5% significance level) a null hypothesis that the odds ratio is one. Using either measurement error pipeline, the confidence intervals are larger than those produced by the TP + GLM pipeline. For the five SNP Analysis TP + GLM had confidence interval width of 0.21, while TP + ME1 had confidence interval width 17.36, and TP + ME2 had confidence interval width of 2.70. Using data from the 10 SNP analyses, all

pipelines had shorter confidence intervals. The TP + GLM pipeline had a confidence interval width of 0.13, TP + ME1 had confidence interval width of 6.58, and TP + ME2 had confidence interval width of 2.15.

	5 SNP Analysis	10 SNP Analysis
Participants	145	377
Clusters	21	46
Mean Cluster Size	7	8.43
Mean Tree Height (Years)	6.86	8.27
Infection Sources	16	43
OR and 95% CI (TP + GLM)	1 [0.89, 1.10]	1 [0.94, 1.07]
OR and 95% CI (TP + ME1)	0.88 [0.04, 17.4]	1.10 [0.18, 6.76]
Mean OR and 95% CI (TP + ME2)	0.80 [0.20, 2.90]	0.96 [0.36, 2.51]

Table 3. Summary of analyses using five and ten SNP cutoffs Summary of analyses using five and ten SNP cutoffs. Mean tree height refers to the height of timed phylogenetic trees in years. TP + GLM refers to a statistical pipeline where infection source labels are first generated using `Transphylo` and then used as response variables in a generalized linear model to calculate an odds ratio. The odds ratio is the odds ratio for the probability of being an infection source given the infection recipient is living without HIV as compared to infection recipients living with HIV.

Discussion

In this study, we used simulations under realistic scenarios for passive TB surveillance to show that the statistical pipeline using `TransPhylo`-generated labels as response variables in a generalized linear model (TP + GLM) had poor frequentist properties for estimating the association between host characteristic and being an infection source under certain conditions. When the difference in probability of infection given contact was high between host categories, as in the 0.3 and 3 settings, TP + GLM had power above 0.95, but due to bias and overly narrow confidence intervals, often failed to capture the true odds ratio. In the case of the 4/7 and 1.75 settings, TP + GLM had low power as well. This poor performance is explained by the sensitivity and specificity of the method used to generate infection source labels. When used in our study setting, the method used in this paper does not accurately label infection sources, leading to substantial under-counts of the true number of infection sources (**Figure A1**). This measurement error results in poor estimation of the odds ratio of interest. The two measurement error models tested as part of a pipeline did not lead to improved coverage or rejection of the null hypothesis.

In all scenarios tested, the bias of using `TransPhylo` labels was such that the estimated odds ratio was always closer to one than the true odds ratio (**Figure 2**). When the true odds ratio was one, the pipeline performed well, with low bias, coverage near 0.95, and the probability of falsely rejecting null hypothesis near 0.05. This suggests that TP + GLM is a conservative pipeline, with bias towards failing to detect a significant effect, unless the effect size is large. In contrast, we found that this pipeline had low probability of false positives for detecting an association. We note that the simulations used in these studies created true timed phylogenies which were used as inputs into `TransPhylo`. In practice, these phylogenies must also be estimated. The errors in reconstruction of phylogenies could further reduce the performance of this pipeline.

Variations of the TP + GLM pipeline have now been implemented in multiple published studies using surveillance data of *M. tuberculosis* [2,29,30]. These studies assumed higher sampling proportions. For example, Xu et. al. study of TB in Spain

used a prior for the sampling proportion with plausible values ranging from 48% to 99% [2]. Depending on the surveillance setting, such high sampling density may indeed be plausible. In our study of TB in Botswana, based on the World Health Organizations estimates of the reporting rate for TB, we assumed a much lower sampling proportion [7]. We recommend carefully considering the sampling scheme of any future studies using surveillance data when conducting research using this pipeline. A high sampling density may be required to reliably estimate effects of interest.

We found that changes to the sampling scheme improved frequentist properties of the TP + GLM pipeline. For example, doubling the number of clusters increased the power of detecting an odds ratio of 1.75 to 0.94 (**Figure 3**). However, just increasing the overall sample size did not improve coverage. While increased sample size decreased variance, and thus confidence interval width, it did not affect bias. In contrast, increasing the sampling density and increasing the sampling window both decreased bias and mean confidence interval width by a small amount. This led to improvements in power, and marginal improvements in coverage (increased sampling density) or only marginal losses in coverage (increased sampling window). The reduction in bias likely follows from improvements in the sensitivity of `TransPhylo` under these settings(**Figure A1**). While the fact that pipeline performance improved when sampling density increased suggests the need for studies with higher sampling density, the improved performance under the increased sampling window suggests there may be room for improvement in the statistical model being used. Currently, `Transphylo` assumes that the sampling proportion is constant throughout the period of the outbreak, while in our simulation scenarios sampling only took place in a fixed number of years of an outbreak (as it often would in a real life surveillance study). Future work could focus on changing this assumption in a `Transphylo`-like model which would better match the the model likelihood to the data generation process.

Applying the `Transphylo`-based data to a population-based TB study in Botswana, we failed to detect an association between HIV status and being an infection source for *M. tuberculosis* transmission. TB is the leading cause of death among people living with HIV [7]. The pathophysiology of HIV-TB co-infection is complicated as HIV-associated immunosuppression reduces *M. tuberculosis* bacterial load in HIV-infected hosts. This observation has led to the hypothesis that people living with HIV have a lower risk of onward TB transmission than people without HIV. [4] Epidemiologic studies investigating this hypothesis have produced conflicting results [30–32]. In light of our simulation study results, our failure to reject the null hypothesis may be due to inadequate statistical power for detecting an association of low to moderate effect sizes when sampling proportion is low. Additional studies using more densely sampled sequences are needed to provider stronger evidence regarding this putative association.

This study had several limitations which should be acknowledged. First, our simulations made several simplifying assumptions which make them imperfect representations of *M. tuberculosis* outbreaks. In particular, we assumed latent and infectious periods were the same for hosts with and without HIV, and further assumed that sampling had no impact on infectiousness, whereas in the real world sampling would presumably be followed by treatment which would impact infectiousness. In our real data analysis, *M. tuberculosis* sequencing data were missing for majority of the enrolled participants. We assumed these data were missing completely at random.

Our findings provide new insights into the use of statistical pipelines that combine pathogen genomic and epidemiologic data to investigate host factors associated with onward transmission of infectious diseases. We recommend careful consideration of potential biases due to imprecise inference of transmission events when using these pipelines. Simulation studies based on the data collection context may improve the interpretation of the results of studies using these pipelines.

Supporting information

Acknowledgements

We thank Xavier Didelot for helping us resolve issues using `TransPhylo` and Paul Bastide and Sebastian Lequime for their help using `nosoi`. We also thank Caroline Colijn, Jessica Stockdale and Vijay Jeevanantham Naidu for sharing their code using `Transphylo`. Research reported in this publication was supported by the National Institute of Allergy And Infectious Diseases of the National Institutes of Health under Award Number R01AI147336 and R01AI097045. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Menardo et. al. The molecular clock of Mycobacterium tuberculosis PLOS Pathogens. 2019 September 12;15(9): e1008067. doi: 10.1371/journal.ppat.1008067
2. Xu et. al. High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain. PLOS Medicine. 2019 October 31;16(10):e1002961. doi: 10.1371/journal.pmed.1002961.
3. Peters et. al. Advances in the understanding of Mycobacterium tuberculosis transmission in HIV-endemic settings. The Lancet Infectious Diseases 2019;19:e65–e76. doi: 10.1016/S1473-3099(18)30477-8.
4. Yates et. al. HIV Infection and the Transmission of Tuberculosis. Journal of Infectious Diseases 2015 May 1;211(9):1510–1510. 10.1093/infdis/jiu628.
5. Gardy et. al. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. New England Journal of Medicine 2011; 364:730–739. 10.1056/NEJMoa1003176.
6. Hall et. al. Using genomics data to reconstruct transmission trees during disease outbreaks. Scientific & Technical Review 2016; 35:287–296.
7. World Health Organization. Global Tuberculosis Report 2020. https://www.who.int/publications-detail-redirect/9789240013131.
8. Zetola NM et. al. Protocol for a population-based molecular epidemiology study of tuberculosis transmission in a high HIV-burden setting: the Botswana Kopanyo study. BMJ Open. 2016 May 9;6(5):e010046. doi: 10.1136/bmjopen-2015-010046.
9. Kohl TA et. al. MTBseq: a comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates. PeerJ. 2018;6:e5895. doi: 10.7717/peerj.5895.
10. Zetola NM et. al. Population-Based Geospatial and Molecular Epidemiologic Study of Tuberculosis Transmission Dynamics, Botswana, 2012–2016. Emerging Infectious Diseases. 2021 March ;27(3):835 - 845.
11. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

12. Stan Development Team. RStan: the R interface to Stan. R package version 2.21.2. 2020.URL:http://mc-stan.org/.

13. Didelot X et. al. Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks Molecular Biology and Evolution. Volume 34, Issue 4, April 2017, Pages 997–1007

14. Lequime S, et. al. nosoi: A stochastic agent-based transmission chain simulation framework in R. Methods in Ecology and Evolution. 2020;11:1002–1007.

15. Beesley LJ and Mukherjee B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. Biometrics. 2020; 1-13.

16. Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS computational biology, 2019 15(4), e1006650.

17. Bouackert, Remco. Ascertainment correction/Sampling bias July 18 2019. https://www.beast2.org/2019/07/18/ascertainment-correction.html

18. Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 2017, 8(1):28-36. doi:10.1111/2041-210X.12628

19. LG Wang, TTY Lam, S Xu, Z Dai, L Zhou, T Feng, P Guo, CW Dunn, BR Jones, T Bradley, H Zhu, Y Guan, Y Jiang, G Yu. treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. Molecular Biology and Evolution 2020, 37(2):599-603. doi: 10.1093/molbev/msz240

20. Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. Biometric (1999), 86, 4, pp. 843-855.

21. Botswana Country Profile https://www.cdc.gov/globalhivtb/where-we-work/botswana/botswana.html Accessed August 8, 2021.

22. C Wymant, M Hall, O Ratmanna, D Bonsall, T Golubchik, M de Cesare, A Gall, M Cornelissen, C Fraser. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity Molecular Biology and Evolution (2018), 35, 3, pp. 719 - 733.

23. EM Volz and SDW Frost. Inferring the Source of Transmission with Phylogenetic Data PLOS Computational Biology (2013), 9(12): e1003397.

24. M Hall, M Woolhouse, A Rambaut. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set PLOS Computational Biology (2015), 11(12), e1004613.

25. F Campbell, A Cori, N Ferguson, T Jombart. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data PLOS Computational Biology (2018), 15(3): e1006930.

26. F Campbell, X Didelot, R Fitzjohn, N Ferguson, A Cori and T Jombart. *outbreaker2*: a modular platform for outbreak reconstruction BMC Bioinformatics(2018) 19(11):363.

27. S Sledzieski, C Zhang, I Mandoiu, MS Bansal. TreeFix-TP: Phylogenetic Error-Correction for Infectious Disease Transmission Network Inference Biocomputing (2021), pp 119-130.

28. SK Gallagher and D Follmann. Branching process models to identify risk factors for infectious disease transmission Journal of Computational and Graphical Statistics (2021), DOI: 10.1080/10618600.2021.2000871.

29. C Yang, L Lu, JL Warren, J Wu, Q Jiang, T Zuo, M Gan, M Liu, Q Liu, K DeRiemer, J Hong, X Shen, C Colijn, X Guo, Q Gao, T Cohen. Internal migration and transmission dynamics of tuberculosis in Shanghai, China: an epidemiological, spatial, genomic analysis Lancet Infectious Disease (2018), 18(7), pp 788-795.

30. B Sobkowiak, L Banda, T Mzembe, AC Crampin, JR Glynn, TG Clark. Bayesian reconstruction of *Mycobacterium tuberculosis* transmission networks in a high incidence area over two decades in Malawi reveals associated risk factors and genomic variants Microbial Genomics (2020), 6, 4.

31. JR Winter, CJ Smith, JA Davidson, MK Lalor, V Delpech, I Abubakar, HR Stagg The impact of HIV infection on tuberculosis transmission in a country with low tuberculosis incidence: a national retrospective study using molecular epidemiology BMC Medicine (2020), 18, 385.

32. K Middelkoop, B Mathema, L Myer, E Shashkina, A Whitelaw, G Kaplan, R Wood, L Becker Transmission of Tuberculosis in a South African Community With a High Prevalence of HIV Infection Journal of Infectious Dseases (2015), 211, 53-61.

33. AC Davison and DV Hinkley. Bootstrap Methods and Their Application, Chapter 5 Cambridge University Press (1997).

34. A Svensson A note on generation times in epidemic models Mathematical Biosciences (2007), 208, 300-311.